Variational Methods for Latent Variable Problems (part 2)

Ryan Giordano (for Johns Hopkins Biostats BLAST working group)

Oct, 2021

Massachusetts Institute of Technology

Outline

Outline for today:

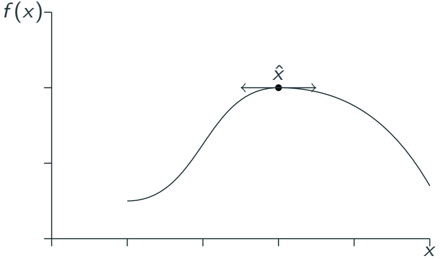
- What counts as variational inference?
- Kullback-Leibler (KL) divergence and “standard” variational inference
- The classical EM algorithm as a special case of variational inference
- Variational inference as a generalization of the EM algorithm
- A quick and incomplete sketch of further topics in variational inference

What counts as variational inference?

Lots of very different procedures go by the name "variational inference." I propose an (idosyncratic) enompassing definition based on the use cases and the name:

Variational inference is inference using optimization.

Think "calculus of variations:" an optimum $\hat{x} = \mathrm{argmax}\ \theta f(x)$ is characterized by $df/dx|_{\hat{x}} = 0$, i.e. where small variations in $\hat{x}$ result in no changes to the value of $f(\hat{x})$.



By this definition,

- The maximum likelihood estimator (MLE) is VI.
- The Laplace approximation to a Bayesian posterior is VI.
- Markov chain Monte Carlo (MCMC) is not VI.

What counts as variational inference?

A more common definition of VI is the following.

Suppose we have a random variable ξ and a distribution $\mathfrak{p}(\xi)$ that we want to know.

Let y denote data and θ a parameter. Examples:

- The variable is θ, and we wish to know the posterior $\mathfrak{p}(\theta|y)$ (Bayes)
- The variable is y, and we wish to know $\mathfrak{p}(y)$ (MLE)
- The variable is y, and we wish to know the map $\theta \mapsto \mathfrak{p}(y|\theta) = \int p(y, z|\theta)dz$ (marginal MLE)

Let $\mathcal{Q}$ be some class of distributions which may or may not contain $\mathfrak{p}(\xi)$.

Variational inference finds the distribution in $\mathcal{Q}$ closest to $\mathfrak{p}$ according to some measure of "divergence" between distributions:

$$\mathfrak{q}^*(\xi) = \operatorname*{argmin}_{\mathfrak{q} \in \mathcal{Q}} D(\mathfrak{q}, \mathfrak{p}).$$

The most common choice of "divergence" is the **Kullback-Leibler** (KL) divergence, though other choices are possible [Li and Turner, 2016, Liu and Wang, 2016, Ambrogioni et al., 2018].

KL divergence

The KL divergence is defined as:

$$\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) := \mathop{\mathbb{E}}_{\mathfrak{q}(\xi)}\left[\log \mathfrak{q}(\xi)\right] - \mathop{\mathbb{E}}_{\mathfrak{q}(\xi)}\left[\log \mathfrak{p}(\xi)\right]$$

Some points to be aware of:

- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) \geq 0$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) = 0 \Rightarrow \mathfrak{p} = \mathfrak{q}$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) \neq \mathrm{KL}\left(\mathfrak{p}||\mathfrak{q}\right)$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right)$ is a “strict” measure of closeness [Gibbs and Su, 2002]

Why use KL divergence?

Phony answer: The KL divergence has an information theoretic interpretation [Kullback and Leibler, 1951].

Real answer: Mathematical convenience (normalizing constants pop out).

Example: the MLE minimizes KL divergence. Suppose that $x_n \overset{iid}{\sim} \mathfrak{p}(\cdot)$, and $\mathfrak{q}(\cdot|\theta) \in \mathcal{Q}$ is a parameteric family of data distributions. Then

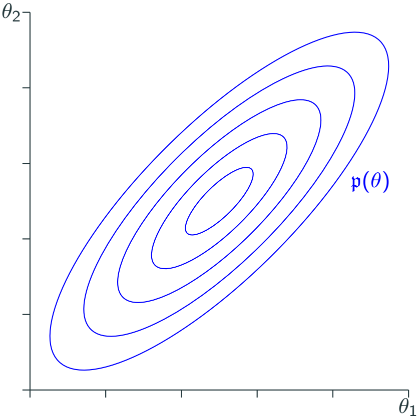
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



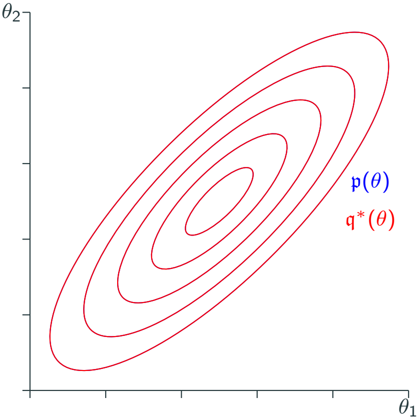
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficiently expressive families recover the target distribution.

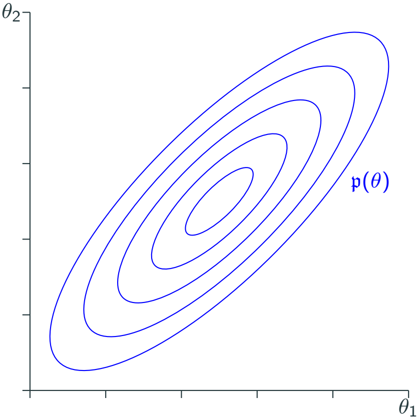
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



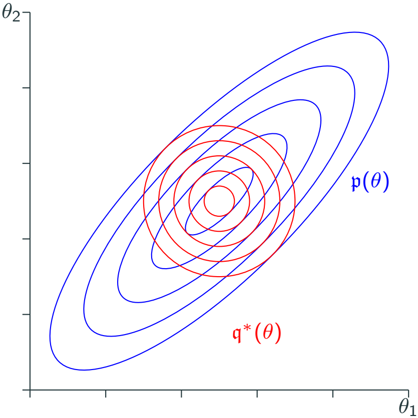
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) =$$
$$-\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



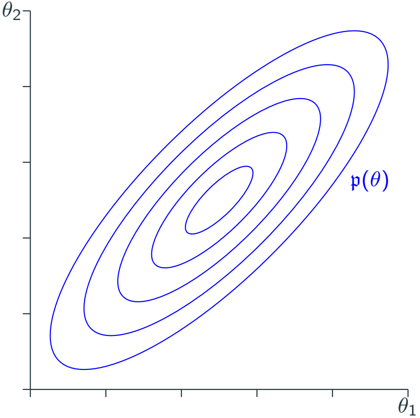
KL minimizers "fit inside" the second argument.

$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) = - \underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right] + \underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?

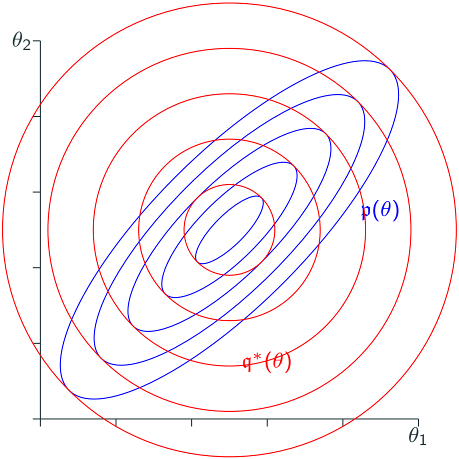


$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}\left[\log \mathfrak{q}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}\left[\log \mathfrak{p}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{Independent bivariate normals}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?



KL minimizers "fit inside" the second argument.

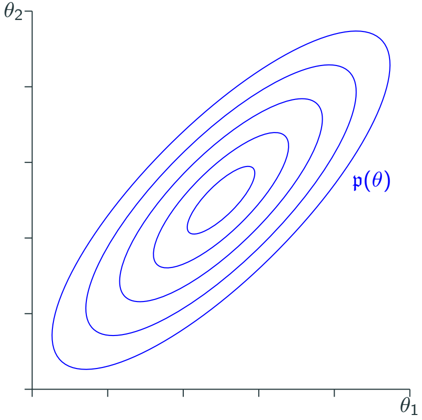
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) =$$
$$- \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$
$\operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \left(- \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)] + \cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]} \right)$?



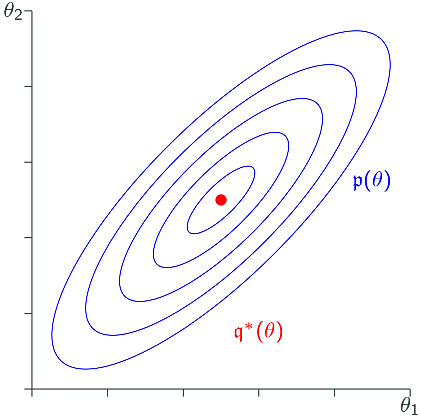
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\left(-\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \cancel{\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]}\right)$?



Without the entropy, the KL minimizer concentrates on the maximum of $\log \mathfrak{p}(\theta)$.

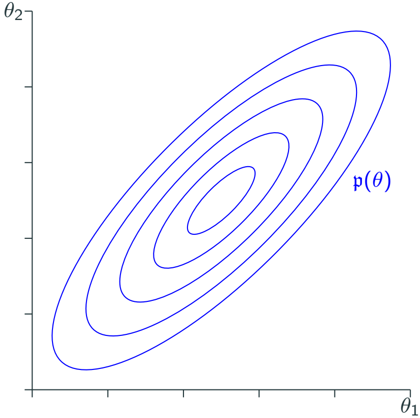
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$
$\underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\left(-\cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right]} + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]\right)$?



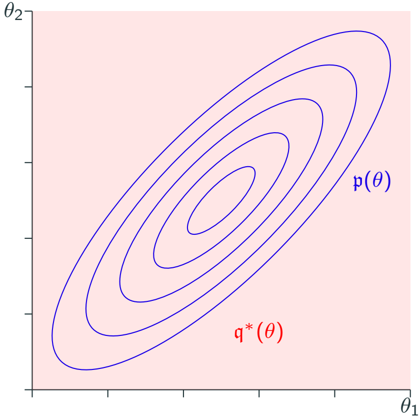
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\left(-\cancel{\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right]} + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]\right)$?



Without $\log \mathfrak{p}(\theta)$, the KL minimizer is infinitely dispersed.

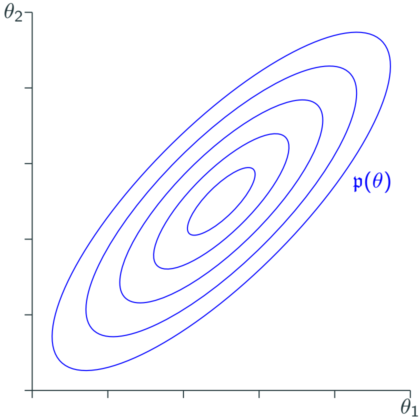
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Point masses}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



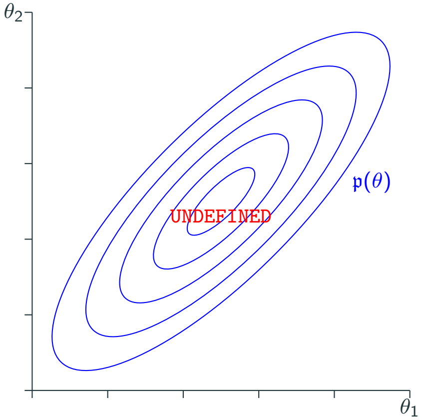
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{Point masses}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q} \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Without a common dominating measure, the KL divergence is undefined.

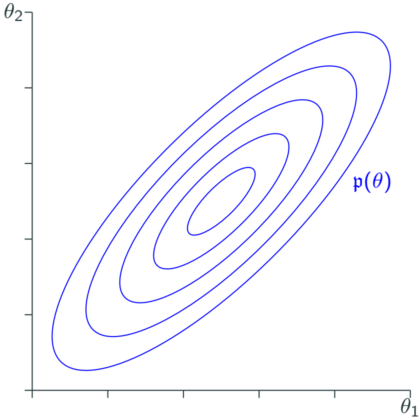
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) =$$
$$- \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q} \in \mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



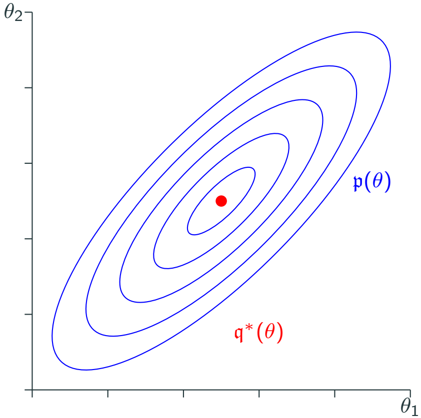
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q} \in \mathcal{Q}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficently concentrated distributions with constant entropy act like a point mass at the maximum of $\log \mathfrak{p}(\theta)$.

Conclusions

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel AJ van Gerven. Wasserstein variational inference. *arXiv preprint arXiv:1805.11284*, 2018.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951.

Yingzhen Li and Richard E Turner. Variational inference with rényi divergence. *stat*, 1050:6, 2016.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.